The IBM LVCSR System Used for 1998 Mandarin Broadcast News Transcription Evaluation

XueFeng Guo, WeiBin Zhu, Qin Shi
IBM China Research Lab
Scott Chen, Ramesh Gopinath
IBM T.J.Watson Research Center

Email: guoxuef@cn.ibm.com
Phone: 8610-6298-6677
Fax: 8610-6298-2439

Abstract

This paper presents the technologies implemented in the IBM's Large Vocabulary Continuous Speech Recognition(LVCSR) system which was designed for 1998 Mandarin broadcast news transcription evaluation task. Compared with the 1997 system, it focuses on acoustic improvements by implementing several new schemes such as LDA and MLLT transformation matrix, BIC model selection criterion, SAT and CAT models. In addition, new language model components and new vocabulary were built. Some other schemes which were tried we also described.

1. System Overview

Speech recognition technology is growing fast and one of recent research focuses has been the transcription of speech data in the real world, such as radio and TV broadcast news(BN). Transcription of broadcast news presents several technical challenges to Large Vocabulary Continuous Speech Recognition(LVCSR) systems. The speech data in broadcast news exhibits a wide variety of speaking styles, environmental and background noise conditions and channel conditions. A typical broadcast news program contains speech and non-speech data from several sources, such as the signature tune of the show, interviews with people on locations possibly under very noisy conditions, interviews over telephone, commercials, etc. This variability exists not only in the English broadcast news, but also in broadcast news material in other languages, such as Mandarin Chinese.

In 1997, IBM developed a basic Mandarin BN transcription system using IBM's LVCSR technologies[1]. It simply performed audio segmentation approach to split the long test data into small segments, and decoded each segment with a "conglomerate" acoustic model trained from all of the acoustic training data provided by Linguistic Data Consortium(LDC), the segments were further clustered by BIC clustering scheme to ensure each cluster is acoustic homogeneous. Finally, iterative unsupervised MLLR adaptation were applied on all the clusters to create the final decoding output.

The 1997 system gave relatively good result which showed the capability of LVCSR system to be used in BN transcription task. In the mean while, it was clear that a lot of work related to acoustic processing could be done to improve the accuracy. In 1998, we focused on signal processing and acoustic modeling, implemented several new schemes into Mandarin HUB4 system and got lots of encouraging results. These new schemes include extending the feature vector from 13 dimensions to 25 dimensions of cepstral, pitch and energy, performing 63-dimensions LDA and MLLT matrix to optimize the feature space, and automatically selecting the number of models using BIC model selecting criterion. These approaches greatly improved the baseline decoding accuracy. Furthermore, we developed Speaker Adaptive Training(SAT) and Cluster Adaptive Training(CAT) models. After baseline decoding, iterative MLLR adaptation was performed like what we did in 1997 evaluation, SAT and CAT MLLR adaptation showed better results than traditional MLLR adaptation. ROVER was tried at the exit of adaptation function but did not bring any improvements, so the scheme was not used in 1998 system.

The paper is organized as follows: section 2 describes the work of signal processing; section 3 explains the new acoustic modeling methods; section 4 introduces the SAT and CAT models; section 5 presents the new language model components and new decoding vocabulary; section 6 shows the adaptation steps; section 7 reports some other schemes we tried for Mandarin evaluation.

2. Signal Processing

In 1997 system, the acoustic front end we used was a 39 dimensions feature space: 11 mel cepstral, pitch and energy were extracted as basic vector, then first and second time derivations were calculated and appended. In 1998, we extended the basic feature vector from 13-dim to 25-dim by adding more cepstral in the vector, this extension gave us limited accuracy improvement. Table 1 shows the comparison results of 13-dim and

25-dim feature vector with the same size of acoustic models.

13-dim vector	25-dim vector
26.1%	25.9%

Table 1. Cepstral extension decoding results

From the number, we may come to a conclusion that the accuracy can not be increased very much by simply adding more cepstral in the feature vector.

When we ran the unsupervised MLLR adaptation in 1997 evaluation, we implemented a simple LDA matrix to convert the feature space into another 39-dimension one, it gave us dramatic improvements at the first iteration. The improvement suggested us that the original 39-dimension space may not represent the Chinese acoustic characteristic very well. In 1998, we made some effort to optimize the feature space, LDA and MLLT are now used.

Linear Discriminate Analysis(LDA) transformation matrix [2], trained from the 30 hours HUB4 acoustic training data, is used to optimize the acoustic feature space. Using the matrix, the input vectors, which are usual feature vectors, spliced together from the adjacent 4 frames on each side of the current frame, are transformed into the new space vector with a smaller dimension. In the 1998 Mandarin HUB4 system, the input vector has 25 dimensions per frame (23 mel cepstral with pitch and energy), and totally 9 frames' vectors are spliced together. We performed the double rotation scheme on the 23 cepstral space to conduct a 60 dimension transformation matrix like in the English system, the pitch dimension of current frame is isolated, and first and second time derivation are computed. These three pitch-based dimensions are attached into the 60-dim matrix to make a 63 dimensions LDA matrix. Using this matrix, the original 9 frames vectors with 25 dimension per frame are transformed to a 63 dimension feature vector to represent the current frame.

Furthermore, based on the 63 dimension LDA matrix, Maximum Likelihood Linear Transformation(MLLT) matrix [4] is extracted from the HUB4 training data, by which the acoustic feature space was transformed from LDA to MLLT space. The final acoustic front end was formed from LDA+MLLT. The idea of MLLT matrix is to find a linear transformation among all the linear transformations to minimize the likelihood loss in the Gaussian models parameters estimation due to lack of training data and computation resource limitation to make the diagonal assumption most valid.

The 63 dimension feature space works very well, both LDA and MLLT gave us significant accuracy improvements. Table 2 shows the improvements of these two schemes. Also, two gender dependent LDA(Gen-LDA) matrices were trained and gave even better decoding results although they are not being used in the evaluation system. We hope CAT could take more advantages of speaker clustering. The front end work brought us the biggest improvement in the 1998 development.

	Small AM	Bigger AM
25dim	25.9%	25.5%
LDA	21.1%	20.6%
LDA+MLLT	20.3%	19.5%
Gen-LDA	19.7%	18.8%

Table 2. Acoustic size and different front end schemes comparison

3. Acoustic Modeling

First experiment we did on acoustic modeling was to try a bigger model based on 1997 experience. In table 2, the small AM has the same size of the acoustic mode that we used in 1997 evaluation which has about 2800 HMM states and 30K Gaussians. A bigger AM is what we tried in 1998. It has around 3000 HMM states and 33K Gaussians. The table 2 shows the bigger AM performed better than the small one. The result suggests an even bigger model.

Bayesian Information Criterion(BIC) was used in 1997 HUB4 system to find the condition change points of the test data, and to cluster segments for unsupervised adaptation. Since BIC is a very effective and commonly used model selection criterion in statistics literature, we expanded the usage to automatically select the number of Gaussian. Unlike the traditional gain threshold method, the BIC does not require pre-defined threshold value which usually comes from experience, it can automatically select the size of models according to the model complexity. The BIC criterion[5] is defined as

$$BIC(M) = \log L(X,M) - \tfrac{\lambda}{2}\#(M) \times \log(n)$$

X is the data set, M is the set of candidates of desired parametric models, $L(X,M)$ is the maximum likelihood function, $\#(M)$ is the number of parameters in model M. By choosing different lambda, we generated different sizes of acoustic models. The decoding results are shown in table 3. All of these models have 4K HMM states The upper row shows the different number of Gaussian.

159K	115K	103K	93K	69K
19.7%	18.5%	18.4%	18.7%	19.0%

Table 3. BIC model decoding results

Besides using BIC criterion to select the model, we change the context dependence from left context only to left and right context to build the decision tree. The advantage of this change is obviously. A particular example is, in Chinese, when two third tone syllables are read together, the tone of former one will be changed to the second tone. In other word, the former syllable's pronunciation is influenced by the right context. Table 4 shows the decoding results comparison of left-only context and left-and-right context. The left-and-right models have 4.4K HMM states.

left		left and right	
159K	19.7%	165K	19.0%
115K	18.5%	107K	18.3%
103K	18.4%	86K	18.4%
93K	18.7%	79K	18.1%
69K	19.0	71K	18.1%

Table 4. BIC model: left vs left and right

The 71K left-and-right model was selected as the final baseline decoding model for 1998 evaluation.

4. SAT and CAT

Speaker Adaptive Training(SAT) scheme has been broadly used in BN transcription task. The idea is to clarify the linguistic acoustic variation from speaker variation. In 1998, we implemented this scheme into the evaluation system, most of the approach is similar as IBM 1998 English evaluation system[3], the only difference is, for Chinese, we use two blocks to process cepstral-based parameters and pitch-based parameters separately. We did not try a single block for both cepstral and pitch, and more study and experiments should be done in the future. The training data we used to train the SAT model is the LDC provided BN acoustic training data clustered by speakers and shows, there are totally 996 training speakers in the training set, A single constrained model space transform was used for each training speaker. The initial model we used is the left and right BIC model which is used as baseline decoding, two iterations of speaker transformation calculation and parameter estimation were performed to train the SAT canonical model.

Cluster Adaptive Training(CAT) is a natural extension of speaker clustering[6]. Unlike the traditional speaker clustering in which the test speaker is clustered into one of the pre-defined speaker groups absolutely, which is also know as “hard” clustering, CAT would like to take advantages of all the speaker clusters by conducting a linear transformation to transform speaker clusters’ mean vector to the test speaker’ s mean vector while keeping the variance and weight vectors unchanged across all the clusters and speakers. The method could be called “soft” clustering. The training and transforming procedures of CAT scheme are quite similar as SAT in practice. The advantages of CAT are, speaker transformation and model estimation become much easier, and training data associated with clusters is greater than with training speakers. We used two clusters(male and female) as cluster set, and iteratively computed the speaker transformation from the two classes and estimated the cluster models to generate the two cluster models.

5. Language Model and Vocabulary

We built single mixture language model as the decoding language model just like what we did in 1997 despite more training data and more components. According to different data smoothing methods, different kinds of language model can be built. In 1998 Mandarin evaluation, we used trigram interpolation LM and maximum entropy(ME) LM.

According to the different levels of importance of the corpus we collected for the 1998 evaluation, we classified the corpus by the content of the corpus and the time of these corpus appeared, used different approaches to build several language models, and mixed them together with different weights. The following are the LM components we used in the HUB4 system.

1. 1997 trigram and ME LMs using 1997 corpus: The corpus includes People's Daily, Beijing Daily and Beijing Evening News, VOA news, and acoustic training transcription.
2. Broadcast news trigram LM using broadcast news corpus: The corpus includes acoustic training transcription, VOA news, China Radio International broadcast news and XinHua News Agency news text provided by LDC in 1997
3. Newspaper trigram LM using newspaper corpus: The corpus includes Marketing Newspaper, People's Daily, Beijing Daily, Beijing Evening News and People's Liberation Army Newspaper purchased from market.
4. Acoustic training transcription trigram and ME LMs using the transcription of acoustic training data.

Among the language model training data, some were distributed by LDC in 1997, while others were purchased from the market or downloaded from Internet,

In order to get better recognition performance, We also expanded the decoding vocabulary to 60K by adding lots of character strings that often appear together as words in Chinese, and some proper names such as names of persons, places and organizations. By building more LM components and expanding the vocabulary, we reduced the perplexity of 1997 evaluation test data to around 300, which is almost half lower than the 1997 system

6. Adaptation

Basically, we keep the adaptation not changed from 1997 system[1]. However, since the initial transcription are required by both SAT and CAT models to compute the transformations of test speakers, and the script's accuracy may influence the property of the transformation, a good initial script is important for the performance of SAT and CAT system. We performed two iterations of MLLR adaptation as well as covariance adaptation on the evaluation data to improve the baseline decoding accuracy, in 1997, only one iteration of mean and covariance adaptation was performed. Then 4 iterations of mean MLLR adaptation were applied on each speaker cluster.

For SAT and CAT, starting from two iterations MLLR and covariance adaptation of baseline model, two iterations MLLR and covariance adaptation are completed, followed by four iterations of MLLR adaptation.

Table 5. shows the results of the 3 adaptation path on the development test data. SAT gives the best result.

MLLR+cov	SAT+MLLR+cov	CAT+MLLR+cov
15.1%	14.3%	14.5%

Table 5. Adaptation results

7. Other Components

Besides the above improvements we made in 1998 evaluation system, there are some other components of the system, such as segmentation, speaker clustering and ROVER. For the segmentation, we made a small modification from the 1997 schemes. The BIC change point detection is run firstly rather than the silence detection, it could gave us more accurate chop points at the condition change points to avoid the influence by the miss segmentation cause by other approaches, table 6 shows this improvement.

1997 Segments	1998 Segments
19.2%	18.9%

Table 6. new segmentation improvements

Speaker clustering was retained from the 1997 system. However, due to SAT and CAT require sufficient amount of adaptation data, and some clusters may contain very little frames, we clustered the speakers by two means, one without frame limitation, and the other with a setting of the lowest frame amount to 5000. We then back-off the small cluster from former clustering groups into their later clustering groups.

We tried ROVER in the development test stages by inputting three scripts generated by three different kinds of system: baseline system, SAT and CAT. The ROVER did not give us any improvements on the accuracy, the reason we guess is both SAT and CAT are developed from the same base model, and the base model is exactly what is used as baseline decoding. Therefore the baseline model, SAT and CAT are quite similar in some sense. The baseline system may include most of the errors of SAT and CAT systems. thus it can not contribute more to the best output selection. Hence, we have to remove the ROVER from our original plan, and submitted the SAT output as the primary result, CAT output was submitted as contrast result. The official character error rate of SAT output is 17.1%, and is 16.9% for CAT. CAT performed even better than SAT in the evaluation test, that is inconsistent with our development test.

8. Conclusions

This paper describes the IBM LVCSR system used in 1998 Mandarin BN transcription evaluation, we implemented several new technologies compared with the 1997 system, the largest gain was obtained from the LDA and MLLT matrix, which shows our current feature space for Chinese speech recognition should and could be optimized through some kinds of transformation; BIC model selection criterion suggests us a model complexity penalty is important to determine the model size; left and right context dependence can work pretty well for Chinese speech recognition; large speaker variance exists in the BN acoustic training set, SAT can efficiently reduce the variance; CAT is much simple than SAT, and "soft" speaker clustering approach has more advantages than the "hard" one.

Acknowledgments

This work is supported by DARPA. Additionally, we would like to thank all the help from speech group of IBM T.J.Watson Research Center and IBM China Research Lab.

References

1. XueFeng Guo, etc. "IBM'S LVCSR SYSTEM FOR TRANSCRIPTION OF BROADCAST NEWS USED IN THE 1997 HUB4 MANDARIN CHINESE EVALUATION", Proceedings of 1998 Broadcast New Transcription and Understanding Workshop.
2. L.R.Bahl, etc. "Robust Methods for using Context Dependent features and models in a continuous speech recognizer", Proc. Intl. Conf. Acoust., Speech and Sig. Proc., 1994
3. Scott Chen, etc. "Recent Improvements to IBM’s Speech Recognition System for Automatic Transcription of Broadcast News", in the same proceedings.
4. R.A. Gopinath, "Constrained Maximum Likelihood Modeling with Gaussian Distributions", Proceedings of 1998 Broadcast New Transcription and Understanding Workshop.
5. Scott Chen, etc. "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition", Proceedings of 1998 ICASSP.
6. M.J.F. Gales, "Cluster Adaptive Training for Speech Recognition", Proceedings of ICSLP98.